

**NUEVO** **EXAMEN**  
**DE ESTADO**

PARA EL INGRESO A LA EDUCACIÓN SUPERIOR

***Cambios para el siglo XXI***

***Sicometria***

*República de Colombia*  
MINISTERIO  
DE EDUCACIÓN NACIONAL



**Revolución educativa en marcha  
¡La educación ya! Asunto de todos**





# Psicometría









# *Sicometria*



## **UNA BREVE HISTORIA DE LA PSICOMETRÍA**

*Son las 5:00 a.m. de 138 a. c. Aún está a tiempo Yu Chen. Sin prisa termina de colocarse su vestido de seda, el mejor que tiene. El hecho de estar despierto desde las 3:00 a.m. no le preocupa mucho puesto que los ejercicios de meditación le hicieron descansar. Ya no era tiempo de preocupaciones, ni de arrepentimientos, todo estaba hecho. Salió de su casa como cientos de miles de personas en toda China que, como él, guardaban esperanzas de contribuir al emperador HAN para que realizara un mejor gobierno en el Imperio.*

*No fue el primero en llegar al sitio indicado. Un rato después ingresaron en orden riguroso de acuerdo con la inscripción que habían realizado previamente y fueron llevados inmediatamente a sus cubículos personales; aquel pequeño lugar que lo albergaría durante 24 horas. Yu Chen se preparó. Colocó la pequeña caja con plumas y tinta encima de la tabla donde escribiría, en donde encontró suficiente papel de arroz para escribir sus respuestas a las preguntas del maestro.*

*Todos esperaban nerviosos el momento de empezar. Todos querían dar lo mejor de sí para apoyar el trabajo de su emperador en el Distrito que fuera necesario. A las 8:00 a.m. el maestro director empezó a dictar la primera pregunta:*

**Describe la producción agrícola de la región, enfatizando las condiciones que la favorecen y los cambios necesarios para garantizar una mejor y mayor producción, con sus consecuencias a corto, mediano y largo plazo, incluyendo las transformaciones administrativas, humanas y técnicas...**

*A Yu Chen no le pareció muy difícil. Le tomaría unas 4 horas el responderla, pero la podría abordar.*

*Luego de escuchar la sexta pregunta (componer un poema a su región) calculó que tendría muy buenas posibilidades de ingresar al servicio civil del emperador y se dio a la tarea de iniciar aquello para lo que se había preparado*





*casi toda la vida.*

Podríamos decir que la educación existe desde el inicio de la humanidad y que la evaluación aparece desde el comienzo de la educación. Por esto no es raro encontrar casos de evaluación educativa en toda la historia de la humanidad y en todas las diversas prácticas educativas que se han desarrollado. Igualmente, la educación se ha venido realizando en diferentes niveles y en cada uno de ellos la evaluación ha adquirido sus propias características que la hacen diferente en cada uno; es así como se puede hablar de la evaluación en el aula de clase como diferente de la realizada en el ámbito institucional.

Es en China de la antigüedad en donde podríamos encontrar la mayor cantidad de ejemplos de evaluación educativa antes de la época moderna. La historia al respecto se remonta a 2200 años A.C.<sup>1</sup> cuando el emperador Chino de la dinastía Shang, hacía que sus funcionarios presentaran pruebas para determinar si eran aptos o no para desempeñarse en el servicio civil. Esos exámenes se refinaron hasta que se introdujeron las pruebas escritas en la dinastía Han (202 a. c. hasta 200 d. c.), fecha en la cual se empezaron a evaluar cinco tópicos:

- ▲ Ley civil
- ▲ Cuestiones militares
- ▲ Agricultura
- ▲ Impuestos
- ▲ Geografía

Para evitar que el calificador hiciera trampa al otorgar mejor resultado a sus conocidos (reconociendo la escritura), las pruebas se reescribían por escribanos contratados especialmente para la ocasión.

La evaluación en China tomó su forma más compleja a partir de 1374 hasta la dinastía Ch'ing (1644 - 1911)<sup>2</sup>, en donde, además de los temas anteriores, se enfatizaba el conocimiento que se tenía sobre las ideas de Confucio. Los candidatos a ocupar cargos públicos

---

<sup>1</sup> DuBois, P.H. (1970). History of Psychological Testing. Boston. Allyn & Bacon.

<sup>2</sup> Greaney, V. Y Kellaghan T. (1995) Equity Issues in Public Examinations in Developing Countries. Washington. World Bank

<sup>3</sup> DuBois P.H. Op. Cit.

pasaban por tres etapas de evaluación<sup>3</sup>:

- ▲ La primera duraba un día y una noche. La persona debía escribir ensayos sobre temas predeterminados y un poema. Pasaban entre el 1% y el 7% de quienes se presentaban.
- ▲ Las evaluaciones del nivel de exámenes de distrito tenían las mismas características de la anterior pero eran más rigurosas y extenuantes (duraban tres días con sus noches). Pasaban entre el 1% y el 10% de quienes alcanzaban este nivel.
- ▲ Los que pasaban las etapas anteriores se presentaban en Pekín a la evaluación final. Sólo pasaba el 3% de quienes se presentaban.

Durante ésta última etapa de evaluación en la china imperial acontecieron ciertos sucesos que resultan curiosos desde una perspectiva como la del actual Examen de Estado colombiano. Dos grandes expertos chinos en evaluación, en el siglo XVII, Ku Yen-wu y Huang Tsung-his, llamaban la atención al hecho que la impresión comercial de libros (apenas en sus comienzos) podría influir negativamente en la educación en el sentido de favorecer la memorización en comparación con otras artes intelectuales; su consecuencia funesta: una menor calidad en las respuestas a los exámenes para ingreso al servicio civil.

Esta suposición se vio confirmada por la gran proliferación de libros relacionados con los exámenes en donde se incluían ejemplos de los ensayos presentados por algunos candidatos. Una práctica común durante estos siglos fue la creación de escuelas especializadas para preparar a los candidatos a presentar los exámenes; las regían intelectuales muy reconocidos pero que no habían tenido éxito en los exámenes. Tal vez el reconocimiento se debía a los libros que escribían sobre cómo interpretar los clásicos de la literatura china de tal forma que pudieran obtener buenos resultados.

Inclusive se presentaron un par de escándalos en los siglos XVI y XVII. En 1595, la persona que obtuvo el segundo lugar había copiado fragmentos de ensayos que aparecían en los libros de preparación para el examen. En 1616, quien ocupó el primer lugar fue descalificado debido a que había copiado totalmente el ensayo de T'ang Pin-yin con el que obtuvo el primer lugar en 1595 y que

apareció en uno de los libros de preparación para el examen.

Pero no todo es malo. Estos exámenes influyeron notablemente en el desarrollo intelectual del imperio Chino y en la divulgación de las ideas a través del material impreso más allá de lo poco que podía imprimir el estado. Benjamín Elman (miembro del consejo de estudios asiáticos de la Universidad de Harvard) argumenta que “los exámenes fueron una obra maestra de la producción social, política y cultural que mantuvieron cohesionada a China durante 400 años” en una de las más importantes dinastías de la historia.

Adicionalmente, el sistema chino de evaluación influyó notablemente en la aparición de pruebas escritas en las escuelas europeas en el siglo XVI y, un par de siglos más tarde, se instituyeron en Europa sistemas de selección a la universidad y al servicio civil, que recordaban las prácticas desarrolladas en China. Estos exámenes públicos han sido una característica esencial de los sistemas educativos europeos desde entonces, quienes llevaron la idea a sus colonias en África, Asia y el Caribe. Hoy en día son una práctica muy frecuente en la mayoría de los países y una fuente rica de información para la transformación de los sistemas educativos.

Estas ideas, en conjunción con la aparición de la psicología experimental, son lo que da la forma actual a la evaluación educativa. En la segunda mitad del siglo XIX, Wilhelm Wundt funda el primer laboratorio de psicología en Leipzig (1879) y experimenta con su **metro de pensamiento** un aparato que pretendía establecer la diferencia entre lo percibido por un sujeto y lo que acontece en la realidad. Sus resultados aportan esencialmente al concepto de medición de las diferencias individuales lo que contribuyó posteriormente al desarrollo del uso de pruebas en psicología para evaluarlas.

Casi de inmediato se desarrolla la medición psicológica conocida como la Epoca de Cobre debido al material del cual estaban hechos la mayoría de instrumentos de medición, cuyos principales exponentes son Sir Francis Galton en Gran Bretaña y James McKeen Cattell en Estados Unidos.

Sir Francis Galton establece, en 1884, el primer laboratorio de psicometría. Allí se encuentran una serie de instrumentos que miden dos aspectos del ser humano, considerados como esenciales por la ciencia de entonces:

### ▲ Aspectos Físicos

Estatura, peso, longitud de la cabeza, longitud de los brazos extendidos, longitud del dedo medio de la mano, longitud del antebrazo, etc.

### ▲ Aspectos Conductuales

Fuerza de la mano para exprimir (medida con un dinamómetro), capacidad vital de los pulmones (medida con un espirómetro), el tono más alto percibido, el tiempo de reacción a estímulos visuales y auditivos, agudeza visual, etc.

Luego, en 1890, James M. Catell, alumno de Galton y dedicado a la medición de diversos aspectos del ser humano, utiliza, por primera vez, las palabras “**Test Mentales**” relacionadas con 10 pruebas que propone aplicar al público en general. Algunas de las pruebas de Catell se mantienen en la tradición de medición de Galton, como por ejemplo, la fuerza requerida para exprimir, el tiempo de reacción a un sonido, el tiempo utilizado para nombrar colores, el grado de presión necesaria para ocasionar dolor<sup>4</sup>, sensibilidad a la diferencia de pesos, entre otras.

A partir de estos trabajos, se presenta un especial desarrollo de la psicometría en estudios de evaluación del ser humano que se originan especialmente en el sistema educativo, que dan comienzo a la nueva época en evaluación conocida como de pruebas (testing). La primera evaluación de este tipo se realiza en Francia en 1905<sup>5</sup>. La Sociedad Libre para el Estudio Psicológico del Niño se encontraba interesada en mejorar la efectividad y eficiencia de las escuelas identificando las causas del fracaso escolar. Desde su perspectiva consideraban que había dos clases de niños que fracasaban: aquellos que podían aprender pero no lo hacían y aquellos que no podían aprender. Alfred Binet, quien se encontraba vinculado con la sociedad mencionada en la época (1904), se dio a la tarea de determinar si el bajo nivel académico de un niño se debía a retardo mental o a alguna otra causa.

---

<sup>4</sup> Galton había mostrado que personas retardadas mentales eran relativamente insensibles al dolor, lo que llevó a Catell a asumir que la sensibilidad al dolor era buen predictor de la inteligencia

<sup>5</sup> Thorndike, R. (1997). The Early History of Intelligence Testing. En Contemporary Intellectual Assessment. New York. The Guilford Press

En 1905, un comité formado por el ministerio público francés, e integrado por Binet y otros miembros de la sociedad mencionada, presentaron la escala Binet-Simon, de 30 tareas cortas, ordenadas por dificultad. El nivel intelectual del niño se definía por la tarea más difícil que pudiera ejecutar correctamente. El cociente intelectual, que introdujo más tarde el propio Binet, relaciona este nivel de habilidad con la edad.

En la prueba de Binet se medían elementos generales del desarrollo del niño a través de un conjunto de tareas heterogéneas. En 1908 y 1911 se hicieron correcciones a la prueba original y se introdujo el concepto de C. I. (Cociente intelectual). En 1916, Lewis Terman, profesor de la universidad de Stanford, hace una revisión a fondo de esta prueba que se conoce actualmente como Stanford-Binet. Su última revisión se realizó en 1986.

El siguiente paso importante en el desarrollo de la evaluación fue la posibilidad de aplicación de pruebas a grupos grandes de personas. Hasta ese momento las pruebas o evaluaciones importadas de Francia, eran esencialmente de aplicación individual o, en algunos casos en que se podían aplicar a grupos, la calificación dependía del juicio del examinador. Con el ingreso de Estados Unidos a la primera guerra mundial se desarrolló a pasos agigantados la posibilidad de aplicación y calificación masiva de pruebas.

En esa época el gobierno de estados unidos encomendó a Robert Yerkes la evaluación de inteligencia de 1.750.000 nuevos reclutas. Yerkes formó un comité que desarrolló los conocidos Army Test (Alpha y Beta). El test Alpha se basó en trabajos no publicados de Otis y consiste en ocho subpruebas:

- ▲ Seguimiento de instrucciones orales
- ▲ Razonamiento aritmético
- ▲ Razonamiento práctico
- ▲ Sinónimos - antónimos
- ▲ Oraciones en desorden
- ▲ Series de números
- ▲ Analogías
- ▲ Información

La noción de pruebas de lápiz y papel, el esquema de evaluación

desarrollado para los Army test, se aplicó, inmediatamente después de la primera guerra mundial, en la industria y en la educación<sup>6</sup>. Se desarrollaron pruebas que se aplicaron a millones de personas durante la década del 20, como el National Intelligence Test. El College Entrance Examination Board, diseña, en 1925, el Scholastic Aptitude Test (SAT) compuesto de preguntas de selección múltiple relacionadas con las oraciones incompletas, analogías y secuencia de números. Esta prueba se administraba a los estudiantes que quisieran ingresar a las universidades en Estados Unidos. La función del Board mencionado fue asumida posteriormente por el Educational Testing Service, entidad que ha diseñado, estandarizado y validado otras pruebas igualmente conocidas<sup>7</sup>.

En 1925 se elaboran pruebas en forma de test con preguntas de selección múltiple, lo que contribuye con el desarrollo de cuestionarios de fácil calificación ocasionando el desarrollo de tecnología que permite la calificación de cuestionarios por medios mecánicos como las máquinas de lectura óptica inventadas en la década del 30. Estos hechos han marcado especialmente el camino que ha tomado la evaluación educativa en el resto de siglo.

En 1957 sucedió un hecho que marcó definitivamente el desarrollo de la educación y de la evaluación educativa, principalmente en Estados Unidos. La Unión de Repúblicas Socialistas Soviéticas (URSS), lanzó al espacio el primer satélite artificial en la historia de la humanidad: el Sputnik I; algunos meses más tarde lanzó al primer ser viviente: la perra Laika y un poco más tarde al primer ser humano: el soviético Yuri Gagarin; en realidad no demoró mucho el viaje de la primera mujer Valentina Teleskova<sup>8</sup>.

Esto llevó a que, de inmediato, se iniciara un cuestionamiento del sector educativo en Estados Unidos, lo que tuvo como consecuencia inmediata la evaluación de sus diferentes componentes: los estudiantes, los currícula, los docentes los programas, la parte administrativa, etc. Se destinaron enormes cantidades de dinero a la evaluación y a la educación y se promulgaron leyes como la Ley

---

<sup>6</sup> Aiken, Lewis. (1996). Test Psicológicos y Evaluación. México D. F Prentice Hall.

<sup>7</sup> DuBois Op. Cit

<sup>8</sup> Wolf, Richard. (1987). Educational Evaluation. En International Journal of Educational Research.

Vol. 11. Londres. Pergamon Press

de Educación Elemental Y Secundaria de 1965, que favorecía el desarrollo de propuestas educativas en los distritos escolares pero sólo si se evaluaban los desarrollos educativos previos. Esto llevó a que los administradores educativos buscaran afanosamente el tipo de evaluaciones educativas que se realizaban en la época para implementarlas lo más pronto posible. Se echó mano del esquema de evaluación con pruebas objetivas.

Toda esta fortaleza que se le dio a la evaluación en educación ocasionó que se convirtiera en una especialidad de la educación desde principios de la década del 60 y que perdiera el sentido que tenía hasta el momento en el contexto educativo: que la evaluación forma parte integral del proceso educativo, sentido que está recuperando en los últimos años.

En la década del 60 se retoman tres perspectivas que unidas se constituyen en el paradigma de la época <sup>9</sup>. El libro de Ralph Tyler (1950) “Principios Básicos del Currículo y la Instrucción”, en el que se propone que la evaluación debe ser parte integral del proceso educativo y debe hacerse para determinar el grado en el cual un programa evaluado promueve el logro de objetivos educacionales, es aceptado rápidamente por muchos administradores educativos debido, principalmente, a su relación con la medición de objetivos observables que surgió a finales de la década del 50.

Estos planteamientos se relacionan con los de Robert Mager (1962) en su libro “Preparando Objetivos para la Instrucción Programada”, en donde se plantean los criterios para la elaboración y formulación de objetivos educacionales. Fue tal el impacto de estos dos libros que muchos evaluadores educativos de principios de los 60, asumieron que la única forma de evaluar un programa educativo, era constatar si se habían logrado los objetivos educacionales planteados en forma medible.

El tercer elemento surge de los planteamientos de B. F Skinner (1958) en su libro “**Máquinas de Enseñanza**”, donde se plantean principios de aprendizaje, establecidos en el laboratorio, para la enseñanza de niños en las escuelas. Esencialmente se plantea la

---

<sup>9</sup> Popham, James. (1987). Two-Plus Decades of Educational Objectives. En: International Journal of Educational Research. Vol. 11. Londres. Pergamon Press.

necesidad de dar material instruccional de manera secuencial, con refuerzos positivos para los aprendices. Los educadores fueron atraídos por la perspectiva de realizar la enseñanza a través de máquinas, debido a la posibilidad de programar la instrucción con materiales que fueran efectivos para los objetivos planteados.

Adicionalmente, se integra, desde el punto de vista de la evaluación, los postulados de Benjamín Bloom (1956) que aparecen en su libro “Taxonomía de los Objetivos Educativos: volumen I, el Dominio Cognitivo”, en donde se plantean las categorías cognitivas relacionadas con el aprendizaje educativo: información, comprensión, aplicación, análisis, síntesis y evaluación.

Estas categorías, en conjunto con la propuesta de Tyler <sup>10</sup> de evaluar los objetivos educativos, forman parte de la cultura evaluativa occidental en el aula de clase y se han constituido en el eje principal de la evaluación escolar.

Dentro de este contexto, y gracias a la inyección económica para el desarrollo de la evaluación educativa, aparecen otros planteamientos que contribuyen a grandes avances en relación con el concepto mismo de evaluación y con los procedimientos para realizarlos. Una de estas miradas es la que plantea Robert Glaser (1963) en su documento “Tecnología Educativa y la Medición del Aprendizaje”. Plantea que la evaluación implica una comparación y que esta comparación puede darse de dos maneras diferentes:

- ▲ **Interindividual.** Los resultados de una persona (o una institución, o programa) se comparan con los resultados de las demás personas que presentan la evaluación con ella. Las inferencias se hacen a partir del porcentaje de personas que se supera.
- ▲ **Intraindividual.** Los resultados de una persona se comparan con lo evaluado. Las inferencias se hacen a partir de lo que la persona puede o no hacer.

El primer tipo de comparación es conocido como Evaluación con Referencia a la Norma y el Segundo como Evaluación con Referencia

---

<sup>10</sup>Tyler es el primero en utilizar el concepto AEvaluación Educativa@, en contraposición con el de Medición Educativa.



a Criterio. Desde ese momento hay una verdadera explosión de pruebas que evalúan con referencia a criterio, puesto que se supone, dan mejor información desde el punto de vista educativo. Sólo había un problema, los modelos matemáticos utilizados para el análisis de datos, no permitían un manejo matemático efectivo de los resultados de las pruebas. Durante muchos años se desarrollaron índices, con problemas ya sea conceptuales o matemáticos, para abordar la problemática planteada por este tipo de evaluación. Sólo a finales de la década del 70 se hace pública la Ítem Response Theory (Teoría Respuesta Pregunta) cuyos modelos matemáticos logit han dado buena cuenta de la Evaluación con Referencia a Criterio.

A finales de la década del 60, Michael Scriven (1967), en su artículo “La Metodología de la Evaluación”, establece la diferencia entre lo que es evaluación sumativa y evaluación formativa. Ante todo plantea que la meta del evaluador es emitir juicios bien informados y que lo esencial del proceso evaluativo es un juicio de valor. Para él, las dos funciones mencionadas se realizan para calcular el valor del objeto luego de estar completamente desarrollado (evaluación sumativa) y para ayudar a desarrollar los objetos (evaluación formativa).

Los anteriores son sólo dos casos que reflejan el impacto de las concepciones de los 60 en la evaluación educativa y en la educación misma y que han guiado el desarrollo de la evaluación luego de su aparición, tanto que aún hoy se reflexiona sobre esos tópicos.

Todas esta discusión se cristaliza en diferentes sistemas de evaluación en diferentes niveles, desde el de aula de clase que adopta la evaluación a través de pruebas objetivas, pero sin la parafernalia técnica en la mayoría de los casos, hasta los sistemas de evaluación nacionales e internacionales.

Dada la necesidad de racionalizar el gasto y la inversión en educación y de redireccionarlo a los elementos que realmente afectan la calidad de la educación, se resalta el papel de la evaluación como generadora de información para orientar los sistemas educativos. Esto hace que no sólo se busque averiguar que tanto sabe un muchacho sino una serie de elementos que puedan estar relacionados con un mayor rendimiento en las pruebas y escudriñar, así, todo el contexto educativo para tratar de entenderlo.

En 1970, la International Association for the Evaluation of Educational Achievement (IEA), realizó la primera evaluación internacional

de ciencias naturales (FISS). Esta evaluación aportó información importantísima en relación con el cumplimiento de las metas educativas en los países participantes, de tal manera que al retroalimentar al sector, se evidenciaron cambios que redundaron en el desarrollo de políticas particulares.

Esta misma organización realizó la segunda evaluación internacional en ciencias entre 1979 y 1988 (cuando presentó los resultados finales), conocida como SISS. Esta segunda evaluación se realizó sobre 57 tópicos de contenido, para cada uno de los cuales se midieron 3 tipos de habilidades:

- ▲ Conocimiento
- ▲ Comprensión de un principio
- ▲ Aplicación de la información y de los principios a la resolución de problemas.

Las pruebas se aplicaron a 3 poblaciones diferentes: 10 años y 4° grado, 14 años y 9° grado y a los estudiantes de último año de educación media. En esta evaluación se utilizan cuestionarios que recogen datos de variables potencialmente asociadas al logro de los estudiantes.

Durante la presente década, la IEA realizó la tercera evaluación internacional (TIMSS) de ciencias naturales incluyendo el área de matemáticas en esta ocasión. Se espera que cada uno de los 41 países participantes profundice en los resultados para lograr los cambios que se propone. Este estudio es de investigación colaborativa entre diferentes instituciones y tiene su centro en The Center for the Study of Testing, Evaluation and Educational Policy del Boston College.

En 1988, el Educational Testing Service (ETS) inicia la primera Evaluación Internacional del Progreso Educativo en matemáticas y ciencias (IAEP I) <sup>11/ 12/ 13</sup>.

---

<sup>11</sup> Es casi una ampliación del National Assessment of Educational Progress que realiza el ETS en Estados Unidos. Inclusive se utilizaron preguntas del Banco de Items en la evaluación Internacional.

<sup>12</sup> Lapointe, A., Mead, N. y Phillips G. (1988) A World of Differences. ETS.

<sup>13</sup> Bertrand, R. Dupuis, F. (1989). A World of Differences. Technical Report. Université Laval.

El enfoque de estas pruebas es esencialmente con referencia a criterio, identificando niveles de habilidad en la población evaluada, por ejemplo, en matemáticas y para la población de 13 años, estos niveles son:

- ▲ Sumas y restas simples
- ▲ Uso de operaciones básicas para resolver problemas
- ▲ Uso de habilidades matemáticas para resolver problemas de dos pasos
- ▲ Comprensión de los conceptos de medición y geometría y resolver problemas más complejos
- ▲ Comprensión y aplicación de conceptos matemáticos más avanzados

En esta primera evaluación participaron 12 poblaciones de 9 países y se utilizaron cuestionarios para recolectar información de factores que potencialmente se relacionan con el logro educativo. En 1994 se realiza la IAEP II con la participación de 24 poblaciones.



xiste un tercer sistema internacional de evaluación educativa el Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (LLECE) cuyos objetivos “consisten en generar estándares regionales, establecer un sistema de información y de diseminación de los avances en relación con ellos, desarrollar un programa de investigaciones sobre las variables asociadas a la calidad de la educación básica y fortalecer la capacidad técnica de los ministerios de educación en el área de evaluación de la calidad de la educación. Además, realizar estudios comparativos sobre calidad de la educación en lenguaje y matemáticas y promover estudios internacionales sobre temas especiales”<sup>14</sup>.

De los tres sistemas internaciones mencionados, los dos primeros se han vinculado, decididamente, con la tercera época de evaluación educativa que involucra en la discusión diversos elementos de diferentes disciplinas que han llevado a la propuesta de evaluación alternativa, planteada como una verdadera alternativa a las prácticas de evaluación en el aula de clase a través, únicamente, de test de rendimiento académico.

En 1989 Robert Glaser<sup>15</sup> propone como conclusión de diversas tendencias de investigación en educación, la evaluación de competencias, entendidas estas como la puesta en práctica en situaciones de la vida cotidiana, de lo aprendido en la escuela. Esta concepción se deriva directamente de las tendencias de la psicología en relación con la cognición humana, especialmente a partir de los trabajos de Howard Gardner relacionados con la inteligencia del ser humano y supera las concepciones de la época (inicios de la década del 80) que relaciona la competencia con la eficiencia y efectividad en relación con los estándares de calidad establecidos<sup>16</sup>.

A partir de este momento y durante toda esta década, se profundiza en el concepto de evaluación educativa y su integración e interacción con todo el contexto educativo y se involucran conceptos como los de evaluación de competencias, evaluación alternativa<sup>17</sup>,

---

<sup>14</sup> UNESCO. (1997) Documentos 1. Marco Conceptual. Santiago Chile

<sup>15</sup> Glaser, R. (1989). New Conceptios of Achievement and Reasoning. En International Journal of Educational Research. Londres. Pergamon Pres

<sup>16</sup> Measuring Competence: defining Performance and Mastery

<sup>17</sup> Worthen, B. (1993). Critical Issues That will Determine the Future of Alternative Assessment. En: Phi Delta Kappan.

evaluación con portafolios, evaluación auténtica, evaluación de ejecuciones, todos relacionados con aspectos cualitativos de la evaluación.

El desarrollo de estas perspectivas implica la vinculación de los elementos cognitivos en la evaluación tales como la resolución de problemas, la creatividad, el pensamiento crítico, entre otros, historia que aún se encuentra en pleno desarrollo.

También existe una historia de los desarrollos técnicos y teóricos relacionados con la medición de atributos del ser humano. En términos generales, la psicometría ha abordado el problema desde dos perspectivas teóricas: la Teoría clásica de los test (mucha de su historia la acabamos de ver) y la Teoría Respuesta al Ítem (IRT). Esta última surge como consecuencia de los problemas que no resolvió la Teoría Clásica de las Pruebas (TCP) como son: el diseño de test, la identificación de preguntas sesgadas o con Funcionamiento Diferencial y el equating (comparabilidad) entre los puntajes de pruebas.

Como dicen Hambleton y Swaminathan (1985) la TCT no provee información acerca del lugar, en la escala de puntajes, de máxima discriminación de los ítems, tampoco ha tenido éxito en detectar el sesgo de las preguntas. Debido a esta y otras razones, en los últimos tiempos se han desarrollado aproximaciones teóricas que sean más apropiadas para solucionar los problemas en la medición de atributos del ser humano.

El propósito de cualquier teoría de la medición es “describir la forma en la cual pueden hacerse inferencias, a partir de las respuestas de una persona a unas preguntas, de características no observables de los examinados, o rasgos medidos por un test”<sup>18</sup>. Las propuestas desarrolladas en el presente se agrupan alrededor de la Teoría Respuesta al Ítem TRI, como se conoce hoy en día.

La historia de esta teoría y sus planteamientos no es tan reciente. Ya para 1936 se planteaban algunos conceptos como el de parámetro y se establecían relaciones con la TCT. Un empujón bastante fuerte a estos modelos lo dio Frederick Lord en 1952 en su tesis de doctorado donde describe el modelo de ojiva normal de dos parámetros, seguido por la propuesta de Birbaum en el sentido de

---

<sup>18</sup> Hambleton y Swaminathan (1985). Item Response Theory. Kluwer. Boston.

sustituir los modelos logísticos por los de la ojiva normal con su correspondiente tratamiento estadístico.

En 1960 Georg Rasch, un matemático Danés, da otro giro a la problemática general de la medición educativa al plantear su modelo logístico de un parámetro o modelo simple., que aparece en el libro *Probabilistic Models for some Intelligence and Attainment Test*. Su trabajo influye en diversas personas relacionadas con la psicometría quienes lo desarrollan y popularizan, entre quienes encontramos a Wright en Estados Unidos, Andrich en Australia, Andersen en Europa, Choppin en Inglaterra.

Durante la década del los 70, se lleva a cabo la gran divulgación de estos modelos. Su dificultad principal radica en la complejidad de los modelos matemáticos empleados y en la imposibilidad de desarrollarlos manualmente, a diferencia de los de la TCP. Es por esto, quizás, que debiera esperarse a la aparición de los computadores personales para que en realidad se difundieran a muchas personas. Instituciones como el ETS (Educational Testing Service) contribuyen a la reflexión sobre el uso de esta teoría de forma amplia.

Es indudable la contribución que los manejos conceptuales y matemáticos de la IRT han tenido en la medición educativa, especialmente y en la medición psicológica en general. La literatura disponible en la actualidad es bastante amplia y el software especializado para los procesos estadísticos es bastante preciso, de muy buena calidad y funcional. En los diversos estudios internacionales como el TIMMS o el IAEP, el estudio de Cívica y Democracia de la IEA y en evaluaciones realizadas por diferentes países como el NAEP en Estados Unidos, las pruebas de logro cognitivo de Colombia, el SIMECAL en Bolivia, el SIMCE en Chile, las evaluaciones nacionales en Australia, las evaluaciones de Holanda, Dinamarca y muchas otras, se utiliza decididamente algún modelo de la IRT para el análisis de ítem, el análisis de prueba, la generación de escalas de calificación y los sistemas de resultados específicos, lo que ha permitido abordar situaciones de evaluación mucho más complejas que las que puede resolver la TCP, como la integración de procesos cualitativos a la medición.

Aun hay mucho que decir sobre las posibilidades de uso de la medición y evaluación educativas y sobre las perspectivas que se presentan hacia el futuro. Esa historia aún se está escribiendo.

## TEORIA CLASICA DE LAS PRUEBAS (TCP)



ómo se ha mencionado, existen dos marcos psicométricos:

- ☐ Teoría Clásica de los Test (**TCT**)
- ☐ Teoría Respuesta al Ítem (**TRI**)

La TCP asume que se pueden construir formas paralelas de una prueba<sup>19</sup>, esto es que midan y evalúen de la misma forma. Se considera que, aunque las formas tengan diferentes preguntas, el hecho de hacer semejantes los índices de dificultad y discriminación de los ítems, garantiza una medición igual con ambas formas (confiabilidad), lo que no permite concebir que una prueba tenga más de una confiabilidad.

Desde el punto de vista de la Teoría de la Generalizabilidad existe un universo de ítems a partir del cual, con procedimientos aleatorios, se pueden seleccionar diferentes grupos de ellos, lo que hace posible que las diferentes formas tengan diferentes confiabilidades. Es más, para una misma situación de prueba, los puntajes pueden tener muchos índices de generalizabilidad dependiendo de los factores que afectan el proceso de medición.

La TCP, que es una forma especial de la teoría de la generalizabilidad en la cual las preguntas de una prueba no se seleccionan de un universo sino que se construyen para ajustarse a él, insiste en la posibilidad de formas paralelas lo que le impide acomodar la noción de que una prueba tenga más de un índice de generalizabilidad: la confiabilidad.

Uno de los problemas centrales que aborda la TCP<sup>20</sup> se refiere a la estimación del puntaje en el universo de preguntas. Se trata de establecer el puntaje de una persona como si hubiera respondido al universo total de preguntas. Como este universo es infinito, es necesario hacer una estimación de ese puntaje, el cual tendrá cierta cantidad de error.

El desempeño de una persona en un conjunto de ítems de una

---

<sup>19</sup> Isaac Bejar. (1983). Achievement Testing. Recent Advances. SAGE. Newbury Park.

<sup>20</sup> Ibidem

prueba puede observarse a partir del puntaje, que es la suma de los puntajes en cada uno de los ítems individuales<sup>21</sup>. La proporción de respuestas correctas, si las preguntas se han obtenido aleatoriamente del universo, es un estimador insesgado de la proporción de preguntas en el universo que una persona puede responder correctamente. De esta forma sería posible conocer el rendimiento de una persona en una disciplina particular, medido a través de una prueba, a partir de las respuestas que dé a cualquier conjunto de ítems (la prueba) obtenido aleatoriamente del universo de preguntas. Esperaríamos que el resultado en cualquiera de estos conjuntos fuera semejante. Aquí lo importante es la validez de contenido. Pero este no es el caso en la TCT, debido a que los ítems se construyen de acuerdo con intenciones particulares.

Estas mediciones no pueden hacerse sin algún error que proviene de diferentes fuentes, como las variaciones propias de las condiciones de aplicación de pruebas, las diferencias en las formas de las pruebas, las variaciones en las ejecuciones de los estudiantes, y otros factores desconocidos.

Supongamos que aplicamos una prueba repetidamente a una misma persona (pensemos que las mediciones son independientes unas de otras y que son idénticas; esto es que la estructura probabilística del experimento no cambia de una aplicación a otra). Podríamos decir que el puntaje de la persona en las diferentes aplicaciones corresponde a su “puntaje observado”, mientras que el valor esperado, calculado a partir de estas observaciones lo llamaremos “puntaje verdadero”. El error correspondería a la diferencia entre el puntaje observado y el puntaje verdadero<sup>22</sup>.

$$e_a = x_a - t_a$$

La variable error tiene una varianza a la que se le denomina **“varianza de error para el examinado a”**,

---

<sup>21</sup> Lord, F Y Novick, M (1968). *Statistical Theories of Mental Test Scores*. Addison-Wesley. Massachusetts.

<sup>22</sup> “La tradición psicométrica ha nombrado a esta entidad como “puntaje verdadero” aunque no posee ninguna propiedad empírica o teórica que sugiera esta terminología. Es un valor esperado, nada más o menos”. Novick, M y Jackson, P (1974). *Statistical Methods for Educational and Psychological Research*. McGraw-Hill. New York.



$$s^2(E_a)$$

cuya raíz cuadrada corresponde al **“error estándar de medición para el examinado a”**

$$s(E_a)$$

que es una medida de la variabilidad de la distribución.

Como

$$t_a$$

es una constante, tenemos que

$$s(X_a) = s(E_a)$$

esto es que el error de medición es un índice de la exactitud en la medida. Si el error estándar de medición tiene una magnitud pequeña significa que el puntaje observado es muy semejante al puntaje verdadero. Debemos recordar que el significado de la magnitud se encuentra en relación con la escala de medición utilizada.

La mayoría de las discusiones en la teoría de las pruebas, se centran en la distribución de resultados de un conjunto de personas y no de un solo individuo, como en el caso que acabamos de presentar.

Dada una población de examinados, cada uno de los cuales tiene un puntaje verdadero, se puede definir una variable  $T$  asociada con la distribución de dichos puntajes. Podemos asumir que ésta distribución tiene un cierto promedio y una cierta varianza. De forma similar podemos considerar los puntajes de error  $E$ , cuyo promedio será cero (0), ya que el promedio de los errores de cada persona es cero. Finalmente podemos considerar los puntajes observados  $X$ , el promedio de su distribución será igual al promedio de los puntajes de los examinados.

De lo anterior se sigue que:

$$X = T + E$$

Que corresponde a la ecuación básica de la Teoría Clásica de las Pruebas (TCP): el puntaje observado es igual al puntaje verdadero más el puntaje de error.

Lo anterior nos lleva a plantear los supuestos básicos de la TCP<sup>23</sup>:

1. El puntaje observado es igual al puntaje verdadero más el puntaje de error.
2. El valor esperado del puntaje de error es cero.
3. La correlación entre los puntajes de error y verdadero es cero.
4. El promedio del puntaje verdadero es igual al promedio del puntaje observado
5. La varianza del puntaje observado es igual a la varianza del puntaje verdadero más la varianza del error.
6. La regresión del puntaje de error en el puntaje verdadero es lineal y con valor constante de cero.

Todos estos supuestos tienen relevancia sólo en el contexto de una población específica ya que ésta se constituye en un evento condicionante.

## **DEBILIDADES DE LA TCP**



xisten diversos aspectos de la TCP que pueden considerarse como debilidades en el contexto de la medición educativa y que mencionaremos a continuación.

### **A - Los valores de las estadísticas de ítems y test dependen de la muestra de examinados.**

En el análisis clásico de ítems el interés se centra en la descripción estadística de cada uno de ellos, considerados como las unidades con las cuales se construye una prueba. El requerimiento básico para un parámetro de ítem es que tenga una relación definida a alguna

---

<sup>23</sup> Ibídem

característica del puntaje total<sup>24</sup>. Adicionalmente, las características estadísticas de un test considerado en su conjunto dependen de las características estadísticas particulares de cada ítem, lo que nos permitirá construir pruebas con propiedades de medición conocidas y en algunos casos óptimas.

Como se colige de lo anterior, es importante, en el diseño de pruebas, contar con estadísticas que describan los ítems de tal forma que se puedan elegir los mejores para conformar la prueba definitiva. Estos estadísticos se obtienen a partir de aplicaciones piloto o de ensayo con muestras representativas de la población a la cual se aplicará la prueba en forma definitiva. Con las palabras “muestras representativas” se alude a grupos de personas en los cuales las estadísticas de los ítems permanecerán invariables. No hay utilidad si en un pilotaje o ensayo se obtienen ciertos estadísticos que no tendrán ninguna relación con la aplicación en la población definitiva.

Los parámetros considerados en la TCP son la dificultad del ítem, su poder de discriminación y su validez<sup>25</sup>.

## DIFICULTAD DEL ÍTEM



Algunos autores utilizan el término de nivel de facilidad de una pregunta en lugar del de dificultad por cuanto permite apreciaciones directas a partir de la lectura del número que la expresa, el cual va de cero (0) a uno (1). El índice de dificultad hace referencia al grado en el cual una población la responde correctamente. Por esto se define a partir de una población de examinados y corresponde a la proporción de individuos que responde correctamente el ítem. Las inferencias se hacen sobre los estimados de dificultad de la pregunta a los parámetros de dificultad para una población bien definida. La inferencia es a un grupo de personas<sup>26</sup>.

Se calcula por medio de:

Donde:

$$P_i = \frac{S U_i}{N}$$

<sup>24</sup> Lord, F y Novick, M (1968). Statistical Theories of Mental Test Scores. Addison-Wesley. Massachusetts.

<sup>25</sup> Ibidem

<sup>26</sup> Hambleton, R. y De Gruitjer, D. Applications of item response models to criterios referenced test item selection. Journal of Educational Measurement. Vol. 20 N° 4.

$P_i$  = índice de dificultad de la pregunta

$U_i$  = respuestas a la pregunta

Si la respuesta es correcta,  $U_i = 1$

Si la respuesta es incorrecta,  $U_i = 0$

$N$  = total de personas que abordan la pregunta

El índice de dificultad es mayor si la muestra tiene habilidad mayor que el promedio de habilidad en la población.

Hoy en día, diversos autores consideran que este no es en realidad un índice de dificultad del ítem sino que es un índice que nos informa sobre la población y por lo tanto lo llaman “**proporción de respuestas correctas**”. Pero como dicen Stenner, Smith y Burduick “**la tradición prevalece**”.

## DISCRIMINACION DEL ITEM



s la varianza, que en una variable dicótoma se puede expresar en términos de la proporción de personas que responde incorrectamente una pregunta <sup>27</sup>. Se calcula a partir de:

$$s_i^2 = p_i q_i$$

Donde:

$p_i$  = índice de dificultad de la pregunta

$q_i = 1 - p_i$

Es claro que la magnitud de la varianza es determinada por cualquiera de los dos valores  $p$  ó  $q$ ; también lo es que el valor máximo es de .25 que ocurre cuando  $p$  y  $q$  valen 0.5 cada una, y este valor disminuye a medida que  $p$  ó  $q$  se desvían de este punto. Al producirse una varianza grande, la incertidumbre respecto a la puntuación verdadera de cualquier persona es mayor.

Como depende tan estrechamente de la proporción de respuestas correctas, si ésta se afecta por la población, la discriminación lo hace en el mismo sentido.

---

<sup>27</sup> Nunnally, J. (1987). Teoría Psicométrica.

## VALIDEZ DEL ÍTEM



Es la correlación biserial puntual entre la respuesta dada a un ítem y el puntaje obtenido en la prueba, excluyendo el aporte de la pregunta al puntaje total en la prueba. Se hacen inferencias a un dominio de contenidos.

Esta correlación es una versión abreviada de la correlación producto momento de Pearson y se usa para especificar el grado de relación que hay entre dos variables, una continua y una dicótoma.

Se calcula a partir de:

$$r_1(t - 1) = \frac{rt_1 St - S_1}{\sqrt{St^2 + S_1^2 - 2S_1 St r_1}}$$

Donde:

$r_{t1}$  = correlación del ítem 1 con la puntuación total, incluido dicho ítem

$St$  = desviación estándar de la prueba

$S_1$  = desviación estándar del ítem

Esta correlación tiende a ser mayor si se estima en una muestra con personas de habilidad heterogénea.

**B - La comparación de los examinados se limita a situaciones en las cuales se les administre el mismo test (o uno paralelo).**

Lord (1968) define pruebas paralelas como aquellas que miden exactamente lo mismo, en la misma escala y que miden con la misma precisión a cada persona. En términos generales la TCP ha definido condiciones estrictas para considerar que dos pruebas son paralelas, las cuales incluyen la igualdad de promedios, varianzas y covarianzas en las diferentes formas<sup>28</sup>. Cuando dos pruebas difieren ligeramente en cualquiera de estos aspectos se hace imposible comparar los resultados de las personas que los abordan.

<sup>28</sup> Embretson, S. 1999. The New Rules of Measurement. Lawrence Erlbaum Associates. New Jersey.

Desde la perspectiva de la TCP, por la razón mencionada anteriormente, no es posible comparar resultados de personas que responden a pruebas con nivel de dificultad diferente.

Ultimamente ha cobrado fuerza la posibilidad de realizar estudios de seguimiento para evaluar el impacto de las transformaciones realizadas en educación (ya sea currículo, prácticas educativas, pedagogía, métodos, etc.). Para ello se requiere recolectar información semejante a lo largo del tiempo y compararla con alguna metodología especial.

En Colombia, el examen de estado vigente a 1999 para ingreso a la educación superior es ideal para realizar un seguimiento puesto que cumple con la condición anterior y, además, es de excelente calidad técnica, de tal manera que se garantizarían resultados significativos. Además, cumple con las condiciones estadísticas mencionadas anteriormente de tal forma que permita la comparabilidad de resultados año tras año y mantener, de esa forma, la equivalencia de los puntajes. No obstante, el esfuerzo que implica desarrollar pruebas con estas características limita posibilidades en otros campos.

Para realizar la comparación año tras año, se requiere del procedimiento de equating (de esta manera se podrían ver los cambios en la población para lo que mide una prueba en particular) lo que permitiría concluir si existen verdaderos cambios en algún aspecto y, si es del caso, encontrar factores asociados con estos cambios. Pero, como se ha mencionado, es difícil, si no imposible en algunos casos, realizar este procedimiento bajo los preceptos de la TCP.

### **C - La confiabilidad, un concepto fundamental, se define en términos de formas paralelas de un test.**

La TCT asume que se pueden construir formas paralelas de una prueba<sup>29</sup>, esto es que midan y evalúen de la misma forma. Se considera que, aunque las formas tengan diferentes preguntas, el hecho de hacer semejantes los índices de dificultad y discriminación de los ítems, garantiza una medición igual con ambas formas

---

<sup>29</sup> Isaac Bejar. (1983). Achievement Testing. Recent Advances. SAGE. Newbury Park.

(confiabilidad), lo cual impide concebir que una prueba tenga más de una confiabilidad.

Teóricamente la confiabilidad se define como la proporción entre las varianzas de la puntuación verdadera y de la puntuación observada

$$r_{xx} = \frac{s_r^2}{s_r^2}$$

Podemos observar que el coeficiente de confiabilidad es cero sólo si la varianza de la puntuación verdadera es cero, lo que subraya la dependencia de este coeficiente de la muestra de donde se obtienen los datos estadísticos; también se puede observar que el valor de 1 (máximo) se alcanza cuando ambas varianzas son iguales.

Como lo menciona Lord y Novick (1968), estos aspectos sugieren que la confiabilidad, en la TCP, es un concepto genérico que se refiere a la precisión en la medición (equivalencia y estabilidad). De ahí se desprenden los diferentes métodos considerados para su estimación<sup>30</sup>:

○ Test – retest. Cuando a una misma muestra de personas se aplica una misma prueba en dos ocasiones diferentes. La estimación se realiza correlacionando los puntajes obtenidos en las dos ocasiones. Por ser la misma prueba, el paralelismo es total. Varios autores plantean que la confiabilidad estimada de esta manera se ve alterado por múltiples fuentes de error que no se tienen en cuenta en la estimación como el aprendizaje de las respuestas entre las dos aplicaciones, la práctica, la memoria, entre otros.

○ Formas paralelas. Cuando se aplican dos formas paralelas (iguales medias y varianzas) a las mismas personas en el mismo momento. De nuevo la confiabilidad corresponde a la correlación entre los puntajes obtenidos. Las diferencias pueden estar ocasionadas por el conocimiento específico de los contenidos de las preguntas, el orden de aplicación de las pruebas, la práctica de una prueba a la otra, etc.

---

<sup>30</sup> Mehrens, W. y Lehmann, I. (1982). Medición y Evaluación en la Educación y en la Psicología. CECSA. México.

○ División por mitades. Es semejante al de formas paralelas. En el presente caso se aplica una sola prueba pero se obtienen puntajes para dos mitades de ella (en la literatura no es claro el procedimiento para dividir la prueba en dos mitades) y luego se correlacionan los puntajes. La dificultad principal radica en el procedimiento de división por mitades.

○ Kuder – Richardson. Es una medida de la consistencia interna de la prueba (denominada así porque la prueba se aplica una sola vez). Es la “correlación media susceptible de obtenerse de todos los posibles cálculos de división en dos mitades”<sup>31</sup>. Depende altamente de la muestra seleccionada para establecer su valor.

○ Coeficiente alfa. Es una generalización de la anterior y produce los mismos resultados. Fue desarrollado por Cronbach en 1951.

**D - Presume que la varianza de los errores de medición es la misma para todos los examinados.**

En este sentido, la Teoría Clásica de las Pruebas asigna un mismo error de medición a todos los puntajes (todas las personas que la abordan). No provee información acerca de la precisión de cada puntaje.

**E - No ha proporcionado soluciones satisfactorias a muchos problemas de medición con pruebas.**

Entre esos problemas se encuentra el diseño de pruebas; la identificación de ítems con funcionamiento diferencial, la comparabilidad de puntajes.

Las pruebas educativas no tienen un sólo y único propósito sino, más bien, apuntan a solucionar diversas problemáticas de acuerdo con las condiciones particulares en donde se presentan. En este sentido, es importante diseñar evaluaciones que respondan a distintas necesidades y que puedan contribuir con información de alta calidad en los diferentes procesos educativos. Esto lleva a pensar que las pruebas se diseñen pensando en estas alternativas y necesidades de tal forma que existan pruebas distintas para diferentes necesidades.

---

<sup>31</sup> Mehrens, W. y Lehmann, I. Op. Cit.



La Teoría Clásica de las Pruebas, considera que la mejor prueba es aquella que tiene una mayor varianza y que discrimina mejor a las personas en un punto particular de la distribución (la media). Se basa en el modelo de la curva normal y por lo tanto exige ciertas particularidades a la distribución de resultados y a los parámetros estadísticos estimados a partir de los puntajes de las personas. Desde este punto de vista, la TCP no puede abordar pruebas que se apliquen a poblaciones de quienes se esperan resultados con distribuciones diferentes a la normal; casos de alta frecuencia en la evaluación educativa.

El funcionamiento diferencial de preguntas (DIF por sus siglas en inglés) hace referencia a que una pregunta pueda ser desventajosa a un grupo cultural específico, en otras palabras, que personas de igual habilidad (con respecto del constructo que mide la pregunta) pero de diferentes culturas muestren o tengan distintas probabilidades de responder la pregunta correctamente, ocasionando una sub o sobre estimación de sus habilidades<sup>32</sup>

Existen muchos métodos que permiten estudiar el funcionamiento diferencial de las preguntas pero, en general, ninguno derivado de la TCP ofrece resultados satisfactorios.

La comparabilidad de resultados en dos aplicaciones diferentes se realiza a través del procedimiento de equating. En algunas ocasiones es necesario que los resultados en exámenes de diverso tipo sean comparables en el tiempo de tal manera que se puedan mantener criterios en ese intervalo, como por ejemplo en los procesos de admisión universitaria. Como se ha mencionado la TCP puede hacerlo pero sólo si se cumplen ciertas características estadísticas de las pruebas.

---

<sup>32</sup> Jonhson, E. (1990). Theoretical justification of the omnibus measure of differential item functioning. IAEF data analysis plan.  
VEALE, J., FOREMAN, D. Assessing cultural bias using foil response data: cultural variation. Journal of Educational Research, 20, pp. 249-258. 1983

# TEORIA RESPUESTA AL ITEM (TRI)

## MODELOS LOGISTICOS

## EN EVALUACION EDUCATIVA



La medición y evaluación educativas se desarrollan a partir de propuestas de la psicometría<sup>33</sup> y toma como fundamentos los de la teoría general de la medición. En este sentido podemos mencionar que la medición de alguna variable, establece la relación que existe entre la variable y la observación que se hace de ella, lo que permite hacer inferencias acerca de esa variable en un sujeto en particular. Adicionalmente el modelo utilizado para relacionar la variable con la observación debe permitir evaluar la validez de la medición.

La medición en educación, no es, en principio, diferente de otras clases de medición, excepto en que no hay mucho consenso en definir cuales son las variables importantes y cuales son las unidades más convenientes para hacer la medición.

Desde esta perspectiva es fundamental la decisión que se tome en relación con la prueba a utilizar y sus particularidades, lo que implica elegir o diseñar la mejor prueba para los propósitos particulares, abordando todos los elementos relacionados con la fundamentación de la prueba tanto en sus componentes teóricos y conceptuales como empíricos, la población a la cual se aplica, la selección de formatos y medios, el establecimiento de criterios de elaboración de preguntas, entre otros. Igualmente, debe decidirse sobre el modelo de medición que se va a utilizar.

En este sentido un modelo de medición “es una función matemática que relaciona la probabilidad de una respuesta correcta a una pregunta con las características de la persona (habilidad) y las características de la pregunta (dificultad)”<sup>34</sup>. Es así como el significado de un resultado en una escala particular está dado por el constructo o marco conceptual seleccionado y no por el modelo en sí.

---

<sup>33</sup> Tyler, Ralph. (1950). Principios básicos del Currículo y la Instrucción. En este libro aparece, por primera vez el concepto de “evaluación educativa”

<sup>34</sup> Stenner y otros, 1983

Por lo tanto, un modelo de medición debe cumplir las siguientes condiciones:

- Una persona con habilidad (en términos psicométricos) alta tiene mayor probabilidad de éxito en un ítem que una persona con habilidad baja.
- Cualquier persona tiene más probabilidad de responder correctamente un ítem fácil que uno difícil <sup>35</sup>

Como consecuencia directa del cumplimiento de estas condiciones, se encuentra que cualquier parámetro (habilidad, dificultad, etc.) debe ser estimado (calculado) independientemente de los demás parámetros. Esto es, que la habilidad de una persona, pueda estimarse independientemente de las preguntas específicas que responda puesto que su habilidad es la “misma” en un momento particular sin importar si responde a una prueba difícil o a una fácil, por ejemplo.

Toda la información acerca de la habilidad de una persona expresada en sus respuestas a un grupo de ítems, se encuentra contenida en la suma simple no ponderada de respuestas correctas a esos ítems. Así, la dificultad de una pregunta puede estimarse independientemente del grupo de personas que la responda.

Estos requisitos permiten formular un modelo matemático y utilizarlo para evaluar qué tan apropiadas son las observaciones para obtener información de la variable en cuestión. No obstante, hay ciertas demandas en relación con el control que se tenga sobre las observaciones: aunque las personas difieran en muchos aspectos, es posible hacer mediciones cuando una dimensión domina la respuesta que se dé a un ítem.

Durante muchos años, en la evaluación educativa se ha utilizado la Teoría Clásica de los Test para diseñar instrumentos, evaluarlos y utilizarlos en diferentes contextos. Entre sus principales inconvenientes, como ya se ha visto, se encuentran la utilización de índices de los ítems que dependen del grupo de personas que los abordan y estimaciones de la habilidad de las personas que

---

<sup>35</sup> Wright, B. Y Mead, R. (1977). Calibrating items and scales with the Rasch model. Research memorandum # 23. University of Chicago.

dependen del grupo particular de preguntas que se incluyen en una prueba<sup>36</sup>.

La psicometría ha avanzado hacia un nuevo sistema de medición: La Items Response Theory o Teoría Respuesta al Ítem (TRI), que tiene dos postulados **a)** la ejecución de una persona en una prueba puede predecirse, explicarse por un conjunto de factores llamados habilidades y **b)** la relación entre la ejecución del examinado y las habilidades que la soportan puede describirse por una función monótonicamente creciente llamada **“función característica del ítem”** o **“curva característica del ítem”** (ICC). Esto último implica que mientras sea mayor la habilidad de una persona, es mayor la probabilidad de responder correctamente una pregunta. Desde esta perspectiva podemos afirmar que la TRI no considera que las formas paralelas sean la justificación de sus resultados (formas que se obtienen aleatoriamente de un universo de preguntas)<sup>37</sup>.

Pueden existir muchos modelos que se diferencian ya sea por la forma matemática de la curva característica del ítem o por el número de parámetros que considera. Todos los modelos tienen por lo menos un parámetro que describe al ítem y por lo menos uno que describe a la persona.

El modelo de respuesta estocástica de Rasch, describe la probabilidad del éxito de una persona en un ítem como una función de la habilidad de la persona y la dificultad de la pregunta, siendo una aproximación estadística al análisis de las respuestas a una prueba y de otros tipos de observación ordinal. Rasch derivó su modelo como una expresión logística simple y demostró que en esta forma los parámetros de la persona y de la pregunta son estadísticamente independientes.

El análisis por el modelo de RASCH construye mediciones lineales de la habilidad de las personas y la dificultad de las preguntas, al mismo tiempo que establece índices de la precisión y exactitud de la medición (ajuste)<sup>38</sup>. Este modelo especifica que cada respuesta útil en una prueba surge de la interacción probabilística lineal entre la

<sup>36</sup> Hambleton, Swaminathan y Rogers, 1991

<sup>37</sup> Bejar, I. (1983). Achievement Testing. Recent advances. SAGE. Newbury Park.

<sup>38</sup> Wright, 1994

medida de la habilidad de una persona y la medida de la dificultad de una pregunta (Rasch, 198). Una forma simple de expresar este modelo es:

$$\log \frac{\text{probabilidad de éxito}}{\text{probabilidad de fracaso}} = \frac{\text{habilidad de la persona}}{\text{dificultad de la pregunta}}$$

El modelo de RASCH presenta las siguientes características<sup>39</sup>:

1. Es matemáticamente simple comparado con otros como el de dos o tres parámetros.
2. Bajo condiciones normales el puntaje bruto de una persona es una estadística suficiente para estimar la habilidad de una persona y el parámetro de las preguntas, lo cual lo hace una extensión de las prácticas actuales en pruebas. Es el único modelo de Respuesta al Ítem (TRI) que es consistente con el puntaje bruto.
3. Predice el comportamiento de preguntas, pruebas y personas con buena efectividad.
4. Establece que la probabilidad de responder una serie de preguntas correctamente está determinada por la habilidad de las personas, esto es que dos personas de igual habilidad tienen la misma probabilidad de responder preguntas fáciles y difíciles (sus curvas características no se cruzan).
5. La probabilidad de responder a la más difícil de dos preguntas debe ser inferior a la probabilidad de responder a la más fácil (las curvas características de las preguntas no deben cruzarse).
6. El problema de la estimación de los parámetros está resuelto.

Desde el punto de vista matemático, el modelo de RASCH es un modelo logístico que se caracteriza “por el supuesto de que

$$\log \frac{P_j}{1 - P_j} = d + b_i x_{ij} + \dots b_k x_{kj}$$

<sup>39</sup> Choppin (1985) y Wright (1977)

Hay varias razones por las cuales es razonable este supuesto. Se supone siempre que la matriz

$$C = \frac{1X_{ij} \dots X_{ki}}{1X_{ij} \dots X_{kj}}$$

tiene rango completo, lo que hace identificable

$$(d, b_i, \dots, b_k)'$$

A partir del logaritmo de la función de máxima verosimilitud

$$\text{se ve que el paso de } p_j \text{ hacia } \log \frac{p_j}{1 - p_j}$$

aparece de una manera natural. Además; significa el paso del intervalo (0,1) hacia la recta real R por medio de la función logit

$$\text{logit}(q) = \log \frac{q}{1 - q}$$

con su inversa

$$\text{logit}^{-1}(X) = \frac{\exp^{(x)}}{1 + \exp^{(x)}}$$

(Weber, 1994) que es una función monotónicamente creciente; equivale a

$$P_r(X_{vi} = B_{vi} | d_i) = \frac{\exp(bv - d_i)}{1 + \exp(bv - d_i)}$$

en el modelo de RASCH (Wright, 1977).

Según Andrich (1988, p.25) otra forma de expresar el modelo es a partir de la función probabilística:

$$P_r\{X_{ni} = 1\} = \frac{L_{ni}}{(1 + L_{ni})} = \frac{\frac{B_n}{D_1}}{G_{ni}} \quad y,$$

$$\Pr \{x_{ni} = 0\} \frac{1}{(1 + L_m)} = \frac{1}{G_{ni}}$$

Donde :

$$G_{ni} = 1 + \frac{B_n}{D_i}$$

es un factor de normalización que asegura que:

$$P_r \{X_{ni} = 1\} + P_r \{x_{ni} = 0\} = 1$$

Además si,

$$B_n < 0 \text{ y } D_i > 0 \Rightarrow 0 < P_r \{X_{ni} = 1\} < 1$$

como se requiere. Esta ecuación es una forma del modelo de RASCH para respuestas dicótomas y es conocida como el modelo logístico simple.

A partir de la estimación de los parámetros se puede calcular la probabilidad de responder correctamente una pregunta y elaborar las ICC (curvas características de las preguntas) que es uno de los supuestos básicos del modelo, con el objeto de establecer si dichas curvas son monótonicamente crecientes (Hamblenton y Cook, 1977; Choppin, 1985).

Adicionalmente al modelo de Rasch o de un parámetro, existen de dos y de tres parámetros, que se basan en principios similares a los planteados anteriormente. Existen algunas pequeñas diferencias en el ámbito conceptual, en términos de los parámetros a partir de los cuales se hace la medición. Una diferencia importante de los modelos de dos y tres parámetros con el de Rasch es que este último es que único que arroja resultados convergentes, esto es que llega a un resultado en la estimación de los parámetros.

Como dice Engelhard G. (1991), la diferencia entre los modelos logísticos (cuyas raíces se encuentran en los trabajos de Catell (1893) y Thorndike (1904)) y la teoría clásica de los test es que estos últimos se han basado en los “puntajes de prueba” mientras que los primeros hacen referencia a “escalas de medición”. En este sentido, la teoría

clásica de los test (que tiene sus raíces en Spearman (1904)), se preocupa por la confiabilidad e inclusive relaciona la objetividad con la forma como se califica una prueba (en otras palabras, la objetividad es un problema de confiabilidad), lo que ocasiona la “paradoja de atenuación” (a medida que un test se hace más confiable, la validez de los resultados medida por la correlación con una variable criterio, se hace más pequeña) como consecuencia negativa.

En la actualidad y debido a los elementos que se mencionan en este documento, se hace notable el uso, cada vez mayor, de los modelos logit para abordar el problema de la medición educativa, de los cuales se encuentran varios que han demostrado ser eficientes en el tratamiento de los problemas que este tipo de medición conlleva.

## ***SUPUESTOS DE LA TRI***



Todo modelo matemático incluye un conjunto de supuestos acerca de los datos en los cuales se aplica y especifica las relaciones entre los constructos descritos en el modelo<sup>40</sup>. En términos generales la TRI considera 3 supuestos básicos:

○ **Dimensionalidad.** La TRI se asume que «un conjunto  $k$  de habilidades soportan la ejecución de un examinado en un conjunto de items. Las  $k$  habilidades definen un espacio  $k$  dimensional, en el cual la localización de cada examinado está determinada por la posición del examinado en cada habilidad. Se dice que el espacio es completo si se han especificado todas las habilidades que influyen en los puntajes de una población de examinados»<sup>41</sup>. Los modelos de dos y tres parámetros requieren de unidimensionalidad en los datos para aplicarlos. Además del modelo de Rasch, hoy en día existen otros modelos de escalamiento multidimensional<sup>42</sup>.

○ **Independencia Local.** Este supuesto es aplicable a diversas posturas en relación con la medición educativa y es que se espera que un estudiante responda a una pregunta en particular sin que recurra a información de otros items para hacerlo correctamente. En otras palabras la ejecución de un estudiante en una pregunta

---

<sup>40</sup> Hambleton, R. y Swaminathan, H. 1985. Item Response Theory: principles and applications. Kluwer. Boston.

<sup>41</sup> Hambleton, R. y Swaminathan, H. 1985. Op. Cit.

<sup>42</sup> Borg, I. Y Groenen, P. 1997. Modern Multidimensional Scaling: theory and applications. Springer. New York.



no debe afectar sus respuestas en otra. Es práctica generalizada en la actualidad el diseñar pruebas en donde conjuntos de ítems dependen de un contexto en particular, del cual dependen las respuestas del examinado; aquí también se aplica la independencia local ya que esta se aplica entre los ítems y no entre ellos y el contexto.

○ **Curvas Características de Ítems.** Es una función matemática que relaciona la probabilidad de éxito en una pregunta con la habilidad medida por el conjunto de ítems que la contienen<sup>43</sup>. Los diferentes modelos de la TRI se diferencian en la forma particular que adquiere la función de probabilidad, la cual incluye el número particular de parámetros del modelo.

## **AJUSTE AL MODELO**



n «modelo de respuesta» (response model) es «una función matemática que relaciona la probabilidad de una respuesta correcta a una pregunta con las características de la persona (habilidad) y las características de la pregunta (dificultad)»<sup>44</sup>. Un modelo no involucra una teoría del constructo a medir y el ajuste de los datos al modelo puede establecerse sin conocimiento de lo que está siendo medido. En este sentido el significado de la escala establecida a partir de las respuestas está dado por el constructo o marco conceptual escogido.

El ajuste de los datos al modelo consiste en que aquellos se encuentren representados adecuadamente por el modelo. Aunque se han planteado diferentes aproximaciones para establecer dicho ajuste<sup>45 46</sup> todos se enfocan hacia la comprobación de los supuestos básicos del modelo o sus implicaciones.

En este sentido podríamos pensar en tres categorías que conformarían esas múltiples miradas de lo que significa verificar el ajuste entre el modelo seleccionado y los datos producidos en una evaluación educativa.

---

<sup>43</sup> Hambleton, R. y Swaminathan, H. 1985. Op. Cit.

<sup>44</sup> STENNER, J. SMITH III, M y BURDICK, D. Toward a theory of construct definition. *Journal of Educational Measurement*, 20, pp. 305-316. 1983

<sup>45</sup> WRIGHT, B. Solving measurement problems with the Rasch Model. *Journal of Educational Measurement*, 14, pp. 97-116. 1977.

<sup>46</sup> CHOPPIN, B. «Bruce Choppin on measurement an education». *Evaluation in Education: An international Review series*, 9, # 1, 1985.

✓ Cumplimiento de los supuestos del modelo por los datos. En el caso de los modelos TRI, los supuestos son los que mencionamos anteriormente: dimensionalidad, independencia local y curvas características de las preguntas.

Existen diversas formas de verificar la dimensionalidad de los datos. Las dos más comunes y que han demostrado buenos resultados son: gráfica de eigenvalores de la matriz de intercorrelaciones inter ítem de tal manera que se pueda determinar un factor determinante y la razón alta entre los dos primeros valores eigenvalores; el otro procedimiento corresponde a la verificación por expertos que siguen un proceso de juicio de jueces particular.

En relación con la independencia local se verifica la estocasticidad estadística interítem.

En relación con las curvas características se verifica que ellas correspondan a la familia de curvas del modelo. El modelo de Rasch exige curvas características que no se entrecruzan.

✓ Cumplimiento de las ventajas derivadas por el uso del modelo. En este caso se hace referencia específica a la particularidad de invarianza de las estimaciones de dificultad y habilidad por el modelo. Esto quiere decir que en cualquier muestra de una misma población los parámetros de ítems y personas permanecen invariantes; es decir que los valores numéricos que se obtienen para la dificultad de las preguntas de una prueba son los mismos en cualquier muestra donde se obtengan. Y el valor del parámetro de las personas (su habilidad) es el mismo en cualquier prueba que respondan (obviamente mediando los procesos de equating necesarios). El modelo de Rasch ha demostrado un mayor cumplimiento de este principio de ajuste, inclusive en estudios realizados en Colombia con los datos de los Exámenes de Estado<sup>47</sup>.

✓ Cercanía entre las predicciones del modelo y los datos. En este sentido se contrastan las distribuciones de las respuestas con las curvas características de cada ítem. De esta manera se puede

---

<sup>47</sup> Pardo, C.; Parra, G, y Grupo de Psicometría. Uso de diversos modelos y procedimientos para el procesamiento de información de los Exámenes de Estado. SNP Bogotá.

reconocer o hipotetizar las posibles causas para comportamientos particulares de grupos poblacionales en relación con los ítems de una prueba.

El modelo de Rasch no supone una distribución determinada de los niveles de dificultad de ítems o de las habilidades de personas. Los procesos de Máxima Verosimilitud utilizados para la estimación de esos parámetros esperan que los errores en las observaciones se distribuyan más o menos normalmente alrededor de los valores esperados. A partir de estas consideraciones es posible calcular dos estadísticas sensibles a respuestas inesperadas que afectan los ítems con dificultades cercanas a las habilidades de las personas (INFIT) o que afectan los ítems con dificultades lejanas a las habilidades de las personas (OUTFIT)<sup>48</sup>.

## FUNCIONAMIENTO DIFERENCIAL DE PREGUNTAS



El funcionamiento Diferencial de Preguntas hace referencia a que un ítem tenga propiedades estadísticas diferentes en grupos poblacionales distintos<sup>49</sup>. Esto es que una pregunta pueda ser desventajosa a un grupo cultural específico, en otras palabras, que personas de igual habilidad (con respecto del constructo que mide la pregunta) pero de diferentes culturas muestren o tengan diferentes probabilidades de responder la pregunta correctamente, ocasionando una sub o sobre estimación de sus habilidades<sup>50 51</sup>.

Existen muchos métodos que permiten estudiar el funcionamiento diferencial de las preguntas pero, en general, se pueden agrupar en dos clases<sup>52</sup>:

<sup>48</sup> Linacre, J. Y Wright, B. A user's guide to Bigsteps Winsteps. Rasch model computer programs. MESA. 1998.

<sup>49</sup> Angoff, W. 1993. Perspectives on Differential Item Functioning Methodology. En: Differential Item Functioning. Lawrence Earlbaum Associates. New Jersey.

<sup>50</sup> JOHNSON, E. Theoretical justification of the omnibus measure of differential item functioning. IAEP data analysis plan. Apéndice 1

<sup>51</sup> VEALE, J., FOREMAN, D. Assessing cultural bias using foil response data: cultural variation. Journal of Educational Research, 20, pp. 249-258. 1983

<sup>52</sup> VAN DER FLIER, H., MELLEBERGH, G., ADER, H., WIJN, M. An iterative item bias detection method. Journal of Educational Measurement, 21, pp. 131-145. 1984.

- ✓ métodos no-condicionales
- ✓ métodos condicionales

Los métodos condicionales, condicionan el sesgo en las preguntas a los niveles de habilidad de quienes abordan la prueba entendidos estos como los puntajes brutos. Las técnicas más utilizadas para verificar el FDP son la de Mantel-Haenzel, la comparación de curvas características de preguntas obtenidas a partir de datos de dos o más poblaciones diferentes y la comparación de los niveles de dificultad obtenidos en dos poblaciones distintas (modelo de Rasch).

## EQUATING



Los resultados que obtienen las personas en una prueba, se utilizan como información que permite tomar diferentes decisiones en distintos niveles. Por ejemplo, en algunos casos las decisiones se toman en el nivel individual como cuando un estudiante decide a qué universidad presentarse para el proceso de admisión. En el nivel institucional los resultados en las pruebas pueden utilizarse para determinar qué personas ingresan a una universidad, por ejemplo. Estos ejemplos corresponden a casos en los cuales las pruebas se administran en múltiples ocasiones, como el Examen de Estado que ocurre en dos ocasiones cada año.

En este último caso mencionado, para el nuevo examen de estado es necesario garantizar que los puntajes obtenidos por los estudiantes en diferentes ocasiones son comparables y que su significado se mantiene.

Esto se consigue con un proceso estadístico denominado *equating* que se utiliza para ajustar puntajes obtenidos con formas diferentes de pruebas de tal manera que estos puntajes se puedan intercambiar<sup>53</sup>, aunque las formas tengan índices de dificultad diferente.

No se debe confundir este procedimiento con otros muy similares como el de comparabilidad de escalas que se utiliza frecuentemente en la TCP, ya que en este caso los puntajes no son intercambiables debido a que se fundamenta en comparaciones de estadísticas a partir de distribuciones de grupos poblacionales..

<sup>53</sup> Kolen, M y Brennan, R. 1995. Test Equating: methods and practices. Springer. New York.

El equating se puede realizar con base en diferentes diseños:

- ✓ Grupos aleatorios. En este caso los examinados se asignan aleatoriamente a las diferentes formas de prueba.
- ✓ Grupo único con contrarrestación. Se aplican las dos formas a un mismo grupo de personas y se contrarresta el efecto del orden de las pruebas al hacer que algunas personas empiecen por una forma y otras por otra.
- ✓ Grupos no equivalentes con ítems comunes. Las diferentes formas de la pruebas comparten ítems comunes o se vinculan a través de una cadena de ítems comunes por segmentos.

## **VALIDEZ**

a cobrado gran fuerza la preocupación por la calidad de las acciones del ser humano en diferentes ámbitos donde se explicitan sus interacciones culturales. Así mismo, se reconoce que la calidad de vida de todo un pueblo está íntimamente ligada con diversos procesos generados por la cultura, siendo el más importante el de la educación.



Podemos reconocer, en la historia, que toda civilización basa su supervivencia y su desarrollo, en general, en las instituciones educativas que genere y, en particular, en los procesos educativos que desarrolle. Nuestra cultura, denominada occidental, ha puesto especial énfasis, desde sus comienzos, en los procesos educativos que desarrolla con todos sus integrantes desde los primeros momentos de vida.

Es evidente la importancia que ha adquirido la educación últimamente, en relación con muchos aspectos de nuestra cultura y en especial con la calidad de vida de las gentes. Es por eso que se han generado diversas herramientas de todo tipo (epistemológicas, metodológicas, etc.), que buscan mejorar los procesos educativos para que respondan a las exigencias del mundo de hoy.

Es muy importante que cualquier sistema educativo sea interactivo con su entorno y pueda reconocer y responder a las necesidades que se le plantean lo que implica la generación de un sistema de evaluación que permita mantener esa interactividad y encauzarla hacia los nuevos requerimientos que se le formulen.

Desde este punto de vista es innegable el valor que cobra la evaluación para contribuir a estos propósitos, orientando la labor que desempeñan los involucrados, interesados y comprometidos con la educación para hacer de este mundo un mejor lugar para vivir.

## UNA PREGUNTA...



uando pensamos en la necesidad de generar un proceso de evaluación en educación (o en cualquier otra área), y nos encontramos en la tarea de diseñar sus elementos específicos, hay una pregunta que ronda nuestras mentes, que orienta la selección o creación de todos estos elementos específicos a integrar en el proceso y que puede plantearse de la siguiente manera.

■ **¿Cómo podemos saber si las inferencias que hacemos a partir de los resultados de un proceso de evaluación, son acertadas?** <sup>54</sup>

La psicometría, que se encarga, entre otras cosas, del diseño de instrumentos<sup>55</sup> de evaluación, ha abordado este tema desde su inicio y ha generado todo un marco de referencia para el análisis de estos instrumentos.

Como ya se puede adivinar, este tópico es el de la validez. En los Standards for Educational and Psychological Testing<sup>56</sup> se incluye resaltado este precepto: **la validez es la consideración más importante en la evaluación de instrumentos**. Más adelante, se puede leer que la validación de una prueba es el proceso de acumular evidencia para argumentar las inferencias que se hagan a partir de los puntajes. Literalmente dice que “aunque la evidencia puede acumularse de muchas formas, la validez siempre se refiere al grado en el cual esa evidencia soporta las inferencias que se hacen a partir de los puntajes. Lo que se valida son las inferencias relacionadas con un uso específico del instrumento, no el instrumento en sí mismo”<sup>57</sup>.

---

<sup>54</sup> Evers, C. (1991). Towards a coherentist theory of validity. International journal of educational research. Vo. 15. Pergamon Pres.

<sup>55</sup> En su sentido más amplio

<sup>56</sup> American Educational Research Association, American Psychological Association y National Council on Measurement and Education. 1985.

<sup>57</sup> Ibídem. Pag 9.

Como se puede observar, la **validez**, es la respuesta a la pregunta planteada desde la epistemología. No obstante, el tema de la validez, desde el punto de vista conceptual, se ha desarrollado de una cierta manera hasta alcanzar los niveles de hoy en día.

## UN POCO DE HISTORIA...



Desde finales de la década del 30, se han explicitado diferentes definiciones de validez, algunas de las cuales aún son de uso común en nuestro medio. Recordemos algunas de ellas. “Validez es el grado en el cual una prueba mide lo que pretende medir”<sup>58</sup>; “la pregunta esencial acerca de la validez de un instrumento, es qué tan bien hace el trabajo para el cual se emplea... En consecuencia validez se define en términos de la correlación entre los puntajes de las personas en la prueba y los puntajes criterio **verdaderos**”<sup>59</sup>; Aún se cita la definición dada por Anastasi (1954) “**validez es lo que la prueba mide y qué tan bien lo hace**”.

No sólo existían muchas definiciones de validez en aquella época, también se promulgaban muchos **tipos** de validez, como la factorial, intrínseca, empírica, lógica y muchas otras. A inicios de la década del 50, se reconocían dos tipos principales, la validez lógica, que incluía el análisis de contenido, los procesos de aplicación de instrumentos, entre otros; y la validez empírica que enfatizaba en el uso del análisis factorial y, especialmente, en las correlaciones entre puntajes y una medida criterio.

Debido a esta gran diversidad de concepciones, la American Psychological Association (APA), combinada con la Association for Applied Psychology, organizó un comité, en 1950, para que estableciera un código de ética que tuviera en cuenta aspectos científicos y aplicados. En 1954 se publica el resultado de su trabajo: el documento *Technical recommendations for psychological test and diagnostic techniques*<sup>60</sup>, en donde aparecen cuatro categorías de validez:

---

<sup>58</sup> Garret, H. E. (1937). *Statistics in psychology and education*. Esta definición la retoma, posteriormente, Cronbach en 1949, aunque le introdujo una pequeña variación: un instrumento es válido en el grado en el cual sabemos lo que mide o lo que predice.

<sup>59</sup> Cureton, E. (1951). *Validity*. En Lindquist (ed) *Educational Measurement*. American Council of Education.

<sup>60</sup> APA (1954).

- Contenido: amplitud y adecuación con que una prueba mide un constructo
- Predictiva: precisión en la predicción de ejecución de una persona
- Concurrente: comparabilidad en la medición de un atributo con dos test diferentes
- Constructo. Es en este documento donde aparece por primera vez.

Para 1966, en los Standards for educational and psychological test and manuals, la validez predictiva y la concurrente se integran en la validez de criterio. En la siguiente publicación de los Standards (1985), se mantienen estas tres categorías.

Como se puede observar hasta el momento, “la concepción tradicional de validez, la divide en tres tipos separados y sustituibles –contenido, criterio y constructo – Esta forma de aproximación es fragmentaria e incompleta, debido especialmente a que no tiene en cuenta la evidencia, tanto de las implicaciones valorativas del significado del puntaje como una base para la acción, como las consecuencias sociales del uso del puntaje”<sup>61</sup>. Esto ocasiona que se genere, desde finales de la década del 80, todo un movimiento que pretende actualizar, reestructurar y profundizar en el concepto de validez.

El nuevo concepto unificado de validez interrelaciona los elementos mencionados anteriormente, “como aspectos fundamentales de una teoría más comprensiva de la validez de constructo que tiene en cuenta tanto el significado del puntaje como la valoración social en la interpretación y uso de la prueba. Esto es, la validez unificada integra consideraciones del contenido, los criterios y las consecuencias en un marco para la comprobación empírica de hipótesis acerca del significado de los puntajes y las relaciones teóricas relevantes, incluyendo aquellas de naturaleza aplicada. Se privilegian seis aspectos distinguibles de la validez de constructo que explicitan la noción de validez unificada. Estos son: aspectos de

---

<sup>61</sup> Messick, S. (1995). Validity of psychological assessment. American psychologist.



contenido, sustantividad, estructura, generalizabilidad, aspectos externos y de consecuencia. En efecto estos seis aspectos funcionan como criterios o estándares para la medición psicológica y educativa”<sup>62</sup>.

## UNA NUEVA CONCEPCIÓN



sí la validez se define como “un juicio evaluativo integral del grado en el cual la evidencia empírica y teórica soportan lo adecuado y apropiado de las interpretaciones y acciones basadas en los puntajes de una prueba u otra forma de evaluación. Validez no es una propiedad de la prueba, es una propiedad del significado de los puntajes de la prueba. Estos puntajes no son sólo una función de las condiciones del ítem o de los estímulos, sino también de las personas que responden y del contexto de la evaluación. Específicamente, lo que debe ser válido es el significado o interpretación del puntaje; lo mismo que cualquier implicación que este puntaje tenga para la acción. La extensión en la cual el significado del puntaje y las implicaciones para la acción se mantienen a través de personas o grupos poblacionales y a través de ambientes o contextos es una pregunta empírica persistente y perenne. Esta es la razón principal por la cual la validez es una propiedad cambiante y la validación es un proceso continuo”<sup>63</sup>.

## CRITERIOS PARA EVALUAR INSTRUMENTOS



omo se mencionó con anterioridad, los aspectos de la validez, funcionan como criterios, a partir de los cuales se puede evaluar, técnicamente, la bondad de un instrumento. El hablar de “validez como un concepto unificado no implica que no pueda diferenciarse en aspectos distintos para subrayar elementos que de otra forma se discutirían superficialmente», tales como las consecuencias sociales de la evaluación de ejecución o el papel que juega el significado del puntaje en un uso aplicado. Lo que se pretende, con esta diferenciación, es proveer con un significado funcional a los aspectos de la validez de tal manera que se puedan desenredar algunas complejidades inherentes a la valoración de lo apropiado, significativo y útil de las inferencias de los puntajes<sup>64</sup>.

---

<sup>62</sup> Messick, S. Op. Cit.

<sup>63</sup> Messick, S. Op. Cit.

<sup>64</sup> Messick, S. Op. Cit.

Los siete aspectos que se privilegian desde esta perspectiva, son:

**1. Contenido.** Incluye evidencia de la relevancia, representatividad y calidad técnica del contenido de la prueba.

◆ La relevancia de las preguntas se evidencia en el enfoque particular de que aparezca en el marco conceptual de la prueba, en donde deben aparecer los criterios particulares para cada una de las variables empleadas. Esto es, que debe proporcionarse la relación entre la variable empleada (y en últimas el ítem utilizado) y la teoría.

◆ La representatividad se observaría a través de la relación entre el énfasis de las variables en el instrumento y el énfasis de elementos correlativos en el marco conceptual.

◆ La calidad técnica del contenido de la prueba, se evidenciaría a través del proceso de análisis de ítem y de prueba. Este análisis involucra aspectos cualitativos y cuantitativos al interior de algún marco de análisis particular. En términos generales existen dos grandes marcos a partir de los cuales se puede realizar este análisis, la Teoría Clásica de las Pruebas y la Teoría Respuesta al Ítem. Estos marcos nos proporcionan evidencia cuantitativa de la calidad del instrumento, ya sea luego de una aplicación experimental o piloto, o luego de una aplicación definitiva. Adicionalmente, es necesario realizar una valoración cualitativa del instrumento<sup>65</sup>.

**2. Esencia.** Argumentos que evidencien las consistencias entre las respuestas al instrumento (evidencia empírica) y los procesos que, desde la teoría, se plantea que asumen los evaluados en las tareas propuestas.

◆ Los argumentos relacionados con este aspecto se encuentran, definitivamente, en el marco conceptual. Es evidente que un marco conceptual puede tomar diversas formas pero, sin importar desde qué perspectiva se elabore, debe contener un acápite en donde se expliciten los procesos que seguirán los evaluados para llegar a

---

<sup>65</sup> La bibliografía relacionada con la verificación de la calidad técnica de los instrumentos de medición y evaluación es abundante. Al respecto se pueden consultar textos como los de Wright, B. (1979). Best test design. Hambleton y Swaminathan. (1985). Item response theory: principles and applications. Entre otros.

las respuestas que se le plantean.

3. Estructura. Se valora la fidelidad que existe entre la estructura de puntuación y la estructura del dominio medido.

◆ Es claro que aquí se analiza el tipo de escala seleccionada para emitir los puntajes en términos de si es la más apropiada para el constructo evaluado. Para esto es necesario seguir el proceso de escalamiento, mediante el cual se asignan números a objetos, de acuerdo con un procedimiento específico. Es necesario aclarar que no todas las formas de asignación de números a objetos llevan, necesariamente, a un escalamiento; diferenciamos, en este sentido, lo que es una **valoración de respuesta** o forma mediante la cual se recogen las respuestas que un grupo de personas da a las preguntas de un instrumento (Ejemplo: acuerdo – desacuerdo; verdadero – falso; etc.), de lo que es una **escala**, la cual resulta de un proceso en donde cada ítem tiene un valor de escala y se refiere a un conjunto de ítems.

◆ En general, una escala tiene como propósito fundamental la generación de puntajes (un solo resultado para un conjunto de ítems)

4. Puntaje: “El término puntaje se utiliza genéricamente en su sentido más amplio para indicar cualquier codificación de consistencias observadas o regularidades en la ejecución de una prueba, cuestionario, procedimiento de observación o cualquier otro instrumento de evaluación (tales como muestras de trabajos, portafolios o simulación de problemas reales. Este uso subsume procedimientos tanto cuantitativos como cualitativos...”<sup>66</sup>.

◆ Adicionalmente se debe tener en cuenta, al construir una escala, su dimensionalidad. Una escala puede tener cualquier número de dimensiones. No obstante, la mayoría de las escalas que se construyen tienen pocas dimensiones, o una sola. Algunos atributos se pueden medir con una sola línea de “números”; por ejemplo la estatura se mide bien sólo con un sistema, el de longitud. Algunos modelos de inteligencia proponen dos grandes dimensiones: verbal y matemática y por lo tanto no se puede medir una sola de ellas para describir la inteligencia de una persona. El modelo de evaluación que se deba utilizar, depende directamente

<sup>66</sup> Messick, Samuel. (1990). Validity of Test Interpretation and Use. ETS.

de la dimensionalidad del atributo o constructo. Esto es, si el atributo es unidimensional, debe utilizarse un método de escalamiento unidimensional y si es multidimensional, utilizar un método de escalamiento multidimensional.

Una preocupación adicional que se debe resolver, de mucha importancia cuando las evaluaciones se enmarcan en un sistema, esto es que se hacen de forma continua durante algún tiempo, se refiere a la comparabilidad de los resultados de tal manera que se pueda hacer algún tipo de seguimiento. Para ello es necesario utilizar un procedimiento de equating particular.

5. Generalizabilidad. Se examina el grado en el cual las propiedades de los puntajes y las inferencias, se generalizan a grupos poblacionales de acuerdo con las necesidades del proceso de medición, de evaluación o de la investigación que se adelante.
6. Aspectos externos. Incluyen evidencia convergente y discriminante a partir de comparaciones multimétodo-multirasgo.
7. Consecuencias. Valora las implicaciones de las interpretaciones de los puntajes como una base para la acción futura; igualmente, las consecuencias presentes y futuras del uso del instrumento especialmente a partir de estudios de sesgos (DIF) y equidad.

Con base en lo anterior se plantean muchos interrogantes que deben resolverse antes, durante y después de diseñada una prueba, que deben guiar todo el proceso de desarrollo de la misma. Tal vez la pregunta que uno puede plantearse hacia el final del proceso es:

**¿Qué argumentos se plantean para las inferencias que aparecen en el análisis de resultados?**

## **RESULTADOS**

En general las medidas educativas se pueden utilizar de dos formas<sup>67</sup>:

- ✓ Las pruebas pueden hacer mediciones referenciadas a la norma, esto es que las ejecuciones de un estudiante son puntuadas e

---

<sup>67</sup> VAN DER LINDEN, W. Criterion-referenced measurement: its main applications, problems and findings. *Evaluation in education*, t, pp. 97-118. 1982.

interpretadas con respecto a las de los demás estudiantes que abordan la prueba. Se hace énfasis en las ejecuciones relativas de los individuos.

✓ Las pruebas pueden hacer mediciones con referencia a un criterio. En este caso se hace énfasis en especificar los referentes (dominios o criterios) que pertenecen a puntajes o puntos específicos a lo largo de un continuo. Se especifica qué tipo de ejecuciones puede realizar un individuo y cual es su repertorio de competencias sin referenciarlo a puntajes o ejecuciones de otros individuos<sup>68</sup>.

El primer caso corresponde a las pruebas en donde se calculan ciertos estadísticos de la población para determinar la escala de resultados. En primer lugar se obtiene el número de respuestas correctas de cada persona que aborda las preguntas. Con este dato se calcula el promedio y la desviación estándar de todos. Se espera que estos datos se distribuyan como la curva normal de tal manera que el promedio quede en la mitad de la distribución de datos y que en total existan 6 desviaciones estándar. Con estos datos se realiza el proceso de estandarización de puntajes y por último se transforma, este resultado, a una escala particular. En el caso del Examen de Estado vigente en 1999, la escala escogida para la presentación de resultados se conoce como T de McAll y tiene un promedio de 50 puntos y una desviación de 10. Como de antemano se sabe que la distribución de puntajes se asemeja a la curva normal, estos puntajes en escala T se pueden interpretar en el mismo sentido.

Por ejemplo si una persona obtiene 60 puntos en la escala T (es decir una desviación estándar por encima del promedio) podríamos decir que esta persona supera (en términos del puntaje) a 84% aproximadamente, del total de personas que abordan la prueba con él. Como se observa, el resultado de alguien se expresa relativamente a los resultados de las demás personas.

En el segundo caso se hace énfasis en la ejecución de las personas que abordan la prueba, independientemente de si se presentan otras personas. Se trata de determinar debilidades, fortalezas; reconocer las ejecuciones particulares de cada uno de tal manera que los

---

<sup>68</sup> HALADYNA, T. y ROID, G. A comparison of two approaches to criterion-referenced test construction. *Journal of Educational Measurement*, 20, pp. 271-282. 1983.

resultados puedan contribuir a los procesos de reorientación personal o institucional. Es decir que un puntaje en particular puede tener significado desde el punto de vista de las disciplinas evaluadas.

El nuevo examen de estado informará significativamente a los usuarios, lo que lo sitúa en el segundo tipo de evaluación educativa.

Lo anterior tiene que ver directamente con el modelo matemático a utilizar. Para el efecto se dispone de dos teorías, la clásica y la de respuesta-pregunta. Las principales ventajas de la segunda son enumeradas por Hambleton y Cook<sup>69</sup> y hacen referencia a que en los modelos de la teoría respuesta-pregunta:

- Es posible comparar examinandos aunque hayan abordado diferentes pruebas que midan el mismo dominio.
- Los parámetros de las preguntas son invariantes aunque se estimen en diferentes muestras de la población. En teoría clásica la calibración de las preguntas son relevantes sólo en el contexto de la muestra donde se realiza
- Proveen una medida de la precisión en la estimación de la habilidad de cada individuo (o cada grupo de individuos con la misma habilidad) mientras que en la teoría clásica sólo se ofrece un único error estándar de medición que se aplica a todos los estudiantes.

De acuerdo con todas estas ventajas se utilizará uno de los modelos de la teoría respuesta-pregunta teniendo en cuenta su aplicabilidad a las pruebas referidas a criterios (las cuales de alguna manera no puede abordar la teoría clásica). Este es el modelo de Rasch o modelo logístico de un parámetro, por razones como las expuestas por Choppin y Wright<sup>70</sup>:

- Es matemáticamente simple comparado con otros como el de dos o tres parámetros.
- Bajo condiciones normales el puntaje bruto de una persona es

<sup>69</sup> HAMBLETON, R., COOK, L. Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement*, 14, pp. 75-96 1977.

<sup>70</sup> Op. Cit.

una estadística suficiente para estimar su habilidad y el parámetro de las preguntas lo cual lo hace una extensión de las prácticas actuales en pruebas. Es el único modelo de respuesta-pregunta que es consistente con el puntaje bruto.

- Predice el comportamiento de preguntas, pruebas y personas con buena efectividad.
- Establece que la probabilidad de responder una serie de preguntas correctamente está determinada por la habilidad de las personas, esto es que dos personas de igual habilidad tienen la misma probabilidad de responder preguntas fáciles y difíciles (sus curvas características no se cruzan).
- La probabilidad de responder a la más difícil de dos preguntas debe ser inferior a la probabilidad de responder a la más fácil (las curvas características de las preguntas no deben cruzarse).
- El problema de la estimación de los parámetros está resuelto. Desde el punto de vista de los algoritmos y procedimientos utilizados en la estimación ofrece resultados convergentes mientras que los de dos o tres parámetros obtienen resultados divergentes en las estimaciones.

## **TIPOS DE RESULTADOS EN EL NUEVO EXAMEN**

on base en estas apreciaciones, se diseñaron 4 tipos de resultados que se pueden obtener a partir de las respuestas de los estudiantes en el Nuevo Examen de Estado.



El examen está compuesto por un núcleo común y un componente flexible. El núcleo común contiene una serie de pruebas de áreas básicas y debe ser abordado por todos los estudiantes. El componente flexible está compuesto por dos líneas: profundización (mayor nivel de complejidad en la evaluación) e interdisciplinar (desenvolvimiento de las personas en distintos escenarios socioculturales). Para cada parte de la estructura se producirán resultados que corresponden a alguno de los siguientes tipos.

## PUNTAJE



Este es un resultado cuantitativo que se obtiene a partir de las estimaciones de la habilidad (en términos psicométricos) de cada estudiante con base en sus respuestas a las pruebas. El resultado *puntaje* se procesa para cada prueba del núcleo común y para la línea interdisciplinar del componente flexible.

La estimación del parámetro habilidad se realiza a través de procesos de máxima verosimilitud que lleva a cabo el software Winsteps<sup>71</sup> que calcula, al mismo tiempo el parámetro de los ítems utilizados (dificultad<sup>72</sup>), el grado de ajuste de los datos al modelo, el equating entre dos formas de la misma prueba, las probabilidades de respuesta para una persona o grupo específico y los estadísticos generales de prueba como los promedios, desviaciones estándar, confiabilidad, entre otros.

En el núcleo común el puntaje (habilidad) indicará la competencia general en cada una de las pruebas tal y como se defina desde la disciplina<sup>73</sup>. En la línea interdisciplinar, el puntaje indicará el desenvolvimiento en los escenarios socioculturales de la prueba seleccionada por el estudiante.

Como se mencionó anteriormente, la interpretación de resultados es con referencia a criterio, es decir que algunos puntajes indicarán el tipo de ejecuciones que puede realizar el individuo. Para ello se realizará un proceso de anclaje de preguntas, es decir que se determina el tipo de competencias para responder a diferentes preguntas en el continuo de la escala de dificultad de ítems.

## RESULTADOS POR TOPICOS



Cada una de las disciplinas evaluadas, al considerar la estructura de las pruebas, puede abordar el problema de clasificar las preguntas de acuerdo con tópicos de interés desde el punto de vista educativo, humano, social, etc., lo que hace posible informar a quien las aborda sobre su desempeño relativo en esos tópicos de tal manera que pueda reorientar sus procesos

<sup>71</sup> Linacre, J y Wright, B. Op. Cit.

<sup>72</sup> La dificultad de un ítem puede entenderse como el grado de exigencia (de la competencia) de una pregunta para ser respondida correctamente por quien la aborde.

<sup>73</sup> Consultar los otros documentos de esta serie



esenciales.

Este desempeño relativo puede ser significativamente superior o inferior al esperado, que podría interpretarse como una fortaleza o debilidad relativa. Se considera que es significativamente superior o inferior, si las diferencias entre su desviación de la media para un grupo de preguntas y su desviación de la media global (C) es superior al doble del error estándar de las diferencias entre estas desviaciones<sup>74</sup>

En el análisis de resultados de la ejecución de los estudiantes, por departamento, en los grupos de preguntas, algunos se citaban como con rendimientos superiores o inferiores al global. En estos casos los departamentos se citaban como desviados de su rendimiento global si las diferencias entre su desviación de la media para un grupo de preguntas y su desviación de la media global (C) es superior al doble del error estándar de las diferencias entre estas desviaciones.

Este tipo de resultado, de tipo cualitativo, se procesará únicamente para las pruebas comunes y se hará a nivel del estudiante, a nivel de la institución educativa (colegio), a nivel departamental y a nivel nacional, de tal manera que se disponga de información que permita cualificar los procesos educativos o las decisiones que se tomen en este ámbito.

## NIVEL DE COMPETENCIA



Para cada una de las pruebas del núcleo común se determinará el nivel de competencia de cada estudiante en cada una de las competencias que mida la prueba.

El nuevo examen tiene como objeto de evaluación las competencias de los estudiantes en contextos disciplinares. Esto implica que se construyan preguntas, en cada prueba, que midan cada una de las competencias descritas en los marcos conceptuales de las pruebas.

El software Winsteps calcula la habilidad de los estudiantes en cada una de las competencias evaluadas en cada prueba, las que se pueden relacionar con niveles particulares de competencia que se establecen con base en el procedimiento de anclaje de preguntas.

<sup>74</sup> Bertrand y Dupuis. 1988. A world of differences. Technical report. ETS. New Jersey.

El resultado es descriptivo (cualitativo) y proporcionará un mapa del desempeño de los estudiantes en las diferentes competencias evaluadas.

## **GRADO DE PROFUNDIZACIÓN**



Este resultado se procesará para la línea de profundización del componente flexible. En esta línea, como se mencionó, se incluyen preguntas de mayor exigencia (mayor dificultad) de tal manera que los resultados puedan ser utilizados como indicadores de fortalezas y contribuyan al proceso de elección de opción profesional.

En cada una de las pruebas de esta línea se podrán reconocer diferentes grados de profundización alcanzados por los estudiantes y que indican el grado de complejidad que pueden abordar y resolver correctamente.

Los grados de profundización se establecen a partir de los niveles de dificultad de las preguntas y existirá una descripción cualitativa del significado, en términos de complejidad, de cada grado. Para cada prueba, se establecerán 3 grados diferentes.